



22883

PATENT TRADEMARK OFFICE

103.1046.01

This application is submitted in the name of the following inventor(s):

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26

Inventor	Citizenship	Residence City and State
Alan ROWE	United Kingdom	San Jose, California

The assignee is Network Appliance, Inc., a California corporation having an office at 495 East Java Drive, Sunnyvale, CA 94089.

Title of Invention

A Mechanism to Survive Server Failures When Using The CIFS Protocol

Background of the Invention

1. *Field of the invention*

This invention relates to transparent recovery of server failures and elective reboots while maintaining consistent data using the CIFS Filesystem protocol.

2. *Related Art*

The Common Internet File system (CIFS) protocol is defined by Microsoft. It enables collaboration on the internet by defining a remote file access protocol that allows applications to share data on local disks and network file servers. CIFS incorporates the same high-performance, multi-user read and write operations, locking, and file-sharing semantics that are the backbone of today's sophisticated enterprise

OFFICE OF THE REGISTER

1 computer networks. With CIFS, users with different platforms and computers can share
2 files without having to install new software.

3
4 CIFS generally runs over TCP/IP, and uses the SMB (Server Message
5 Block) protocol found in Microsoft Windows® for file and printer access; therefore,
6 CIFS will allow all PC applications, not just Web browsers, to open and share files across
7 the Internet.

8
9 With CIFS, both the client and the server maintain state about filenames,
10 file contents, directories, and various other aspects of the files and directories; thus CIFS
11 is a “stateful” protocol. File content is cached via a cooperative process between client
12 and server code, and this is where problems can occur. The state survives only as long as
13 the session between the server and the client survives, and this session survives only as
14 long as the underlying network connection (generally TCP/IP) survives.

15
16 When a server that is currently supporting one or more sessions fails or has
17 to be purposefully rebooted, all sessions being supported are lost. CIFS has no protocol
18 for re-establishment of a session after such a fatal error, or for synchronization of the
19 client/server state to the pre-crash state. CIFS does support fault tolerance in the face of
20 network and server failures where some CIFS clients can restore connections and reopen
21 files that were open prior to interruption, however, any data that was currently being
22 edited that had not been saved is lost. As a result, a server failure is regarded as a
23 catastrophic event in the CIFS world.

24
25 Accordingly, it would be advantageous to provide a technique that
26 addresses reestablishing server client sessions that were utilizing CIFS after a server

1 failure or elective reboot so that operation resumes where it ended prior to server
2 unavailability.

3
4 Summary of the Invention

5
6 The invention includes a method and system for re-establishing sessions
7 between a server and clients that were using the CIFS protocol. Two types of situations
8 may occur. The first type occurs when a system administrator purposefully reboots the
9 server or a purposeful takeover occurs in a clustered configuration. For these elective
10 reboots of a server a series of tasks are performed; (1) the server stops accepting
11 incoming CIFS requests, (2) the server completes processing of active CIFS requests, (3)
12 all active CIFS state and networking state is captured in non-volatile storage (CIFS data
13 structures are static at this point), (4) the server is rebooted, (5) state is rebuilt for the
14 rebooted machine from that which was saved in non-volatile storage; in a takeover
15 configuration state is made available through transmission or some form of non-uniform
16 memory access, and (6) incoming CIFS requests are once again accepted and operation
17 resumes.

18
19 The second type occurs when the server reboots without warning or there is
20 an unplanned takeover due to server failure. These unplanned occurrences require the
21 following tasks be performed; (1) state is saved persistently at predetermined intervals to
22 non-volatile storage, (2) when the system crashes and reboots or is taken over, state is
23 restored from the non-volatile storage, or in a takeover configuration, state is made
24 available through transmission to a subsequent machine or through some form of non-
25 uniform memory access, (3) operations that were in progress resume at the steps they
26

1 were at prior to the crash, (4) new CIFS requests are now accepted. All of the preceding
2 is transparent to the clients and no data are lost.

3 4 5 Brief Description of the Drawings

6
7 Figure 1 illustrates a block diagram of a system for server failure survival
8 when using the CIFS protocol.

9
10 Figure 2 illustrates a file server elective reboot/takeover process in a system for
11 server failure survival when using the CIFS protocol.

12
13 Figure 3 illustrates a file server non-elective reboot/takeover process in a system
14 for server failure survival when using the CIFS protocol.

15
16 Figure 4 illustrates a file server non-elective takeover process in a system to
17 survive server failures when using the CIFS protocol.

18
19 Figure 5 illustrates critical state saving points in a mechanism to survive server
20 failures when using the CIFS protocol.

21 22 Detailed Description of the Preferred Embodiment

23
24 In the following description, a preferred embodiment of the invention is
25 described with regard to preferred process steps and data structures. Embodiment of the
26 invention can be implemented using general purpose processors or special purpose

1 processors operating under program control, or other circuits, adapted to particular
2 process steps and data structures described herein. Implementation of the process steps
3 and data structures described herein would not require undue experimentation or further
4 investigation.

5 6 Lexicography

7
8 The following terms refer to or relate to aspects of the invention as described
9 below. The descriptions of general meanings of these terms are not intended to be
10 limiting, only illustrative.

- 11
- 12 • client and server – These terms refer to a relationship between two devices,
13 particularly to their relationship as a client and server, not necessarily to any particular
14 physical devices.
 - 15
 - 16 • Client device and server device – These terms refer to devices taking on the role of a
17 client device or a server device in a client-server relationship (such as an HTTP web
18 client and web server). There is no particular requirement that any client devices or
19 server devices be individual physical devices. They can each be a single device, a set
20 of cooperating devices, a portion of a device, or some combination thereof.
 - 21
 - 22 • Procedure – A procedure is a self-consistent sequence of computerized steps that lead
23 to a desired result. These steps are defined by one or more computer instructions.
24 These steps are performed by a computer executing the instructions that define the
25 steps. Thus, the term “procedure” can refer to a sequence of instructions, a sequence
26 of instructions organized in a programmed-procedure or programmed-function, or a

1 sequence of instructions organized within programmed-processes executing in one or
2 more computers.

- 3
- 4 • CIFS – Common Internet File System protocol defines a standard for remote file
5 access using millions of computers at a time across different platforms that can share
6 files.
 - 7
 - 8 • NetBIOS – An application programming interface (API) that augments the DOS
9 BIOS by adding special functions for local-area networks (LANs).
 - 10
 - 11 • NBT – An implementation of Netbios over TCP/IP.
 - 12
 - 13 • SMB – Server Message Block. A message format used by DOS and Windows too
14 share files, directories and devices. NetBIOS is based on the SMB format.
 - 15

16 As noted above, these descriptions of general meanings of these terms are not
17 intended to be limiting, only illustrative. Other and further applications of the invention,
18 including extensions of these terms and concepts, would be clear to those of ordinary
19 skill in the art after perusing this application. These other and further applications are
20 part of the scope and spirit of the invention, and would be clear to those of ordinary skill
21 in the art, without further invention or undue experimentation.

22

23 *System Elements*

24

25 Figure 1 shows a block diagram of a mechanism to survive server failures
26 when using the CIFS protocol.

1
2 A preferred embodiment of the system 100 can include a client device 110,
3 a client communication link 120, a communications network 130, a first server 140, a
4 second server 150, a server communications link 160, an interconnect 170, and a mass
5 storage 180.

6
7 The client device 110 includes a processor, memory, mass storage (not
8 shown but understood by one skilled in the art). Typically, the client device 110 is
9 associated with a user.

10
11 The client communication link 120 couples the client device 110 to the
12 communications network 130. In a preferred embodiment, the communications network
13 130 includes an Internet, intranet, extranet, virtual private network, enterprise network, or
14 another form of communication network.

15
16 The first server 140 includes a processor, a main memory (not shown but
17 understood by one skilled in the art), and a first non-volatile storage 141. In a preferred
18 embodiment the first server 130 and the client device 110 are separate devices, however,
19 there is no requirement in any embodiment that they be separate devices. In a preferred
20 embodiment, the first non-volatile storage 141 includes any electronic storage medium
21 capable of retaining state without power or by some auxiliary power source (such as;
22 non-volatile random access memory, magnetic and optical drives).

23
24 The second server 150 includes a processor, a main memory (not shown but
25 understood by one skilled in the art), and a second non-volatile storage 151. In a
26 preferred embodiment the second server 150 and the client device 110 are separate

1 devices, however, there is no requirement in any embodiment that they be separate
2 devices. In a preferred embodiment, the first non-volatile storage 151 includes any
3 electronic storage medium capable of retaining state without power or by some auxiliary
4 power source (such as; non-volatile random access memory, magnetic and optical
5 drives).

6
7 Additionally, the invention is applicable to both a standalone server and a
8 server cluster; however, the second server 150 is used only in applications of the
9 invention where the functions of the first server 130 are to be taken over by the second
10 server 150. There is no requirement in any embodiment that the second server 150 be
11 present in non-takeover applications of the invention.

12
13 A server communications link 160 couples the first server 140 and the
14 second server 150 to the communication network 130.

15
16 An interconnect 170 couples the first server 140 to the second server 150
17 providing bi-directional communication between the two servers.

18
19 The mass storage 180 is coupled to both the first server 140 and the second
20 server 150. In a preferred embodiment the mass storage 180 includes magnetic and
21 optical disk arrays, and other devices capable of storing relatively large amounts of data.

22
23
24 *Method of Operation – Elective Takeover and Elective Reboot*

1 Figure 2 illustrates a file server elective reboot/takeover process, indicated
2 by general reference character 200. The file server elective reboot/takeover process 200
3 initiates at a 'start' terminal 201. The file server elective reboot/takeover process 200
4 continues to a 'boot system' procedure 203 which enables the first server 140 to boot.

5
6 A "flag active" decision procedure 205 determines whether the first server
7 140 is rebooting following an elective reboot. If the 'flag active' decision procedure 205
8 determines that the first server 140 has been subjected to a reboot, the file server elective
9 reboot/takeover process 200 continues to a 'restore state' procedure 227.

10
11 A 'receive CIFS requests' procedure 207 allows user requests to be
12 received by the first server 140.

13
14 A 'process CIFS requests' procedure 209 allows the first server 140 to
15 respond to requests from the client device 110 by providing access to data contained in
16 the mass storage 180.

17
18 An 'initiate elective process?' procedure 211 determines whether the
19 system is to be purposely taken offline (e.g. by the systems operator for maintenance
20 purposes). If the 'initiate elective process?' procedure 211 determines that an elective
21 shutdown has not been initiated, the file server elective reboot/takeover process 200
22 continues to the 'receive CIFS requests' procedure 207.

23
24 An 'ignore CIFS requests' procedure 213 causes the server device 140 to
25 ignore all incoming CIFS requests from the client device 110. This is perceived by the
26 client device 110 as a network delay and will not by itself terminate the session. The

1 client device 110 will resubmit CIFS requests until accepted or until the session is timed
2 out (approximately 45 - 60 seconds from receipt of the first rejection by the client device
3 110) which ever comes first. The invention enables acceptance of CIFS requests prior to
4 a session timing out.

5
6 A 'drain CIFS requests' procedure 215 ensures that all currently active
7 CIFS requests are processed to completion.

8
9 An 'elective takeover?' decision procedure 217 determines whether an
10 elective takeover has been selected by the systems operator. If the 'elective takeover'
11 decision procedure 217 determines that an elective take over has been selected by the
12 systems operator the file server elective reboot/ takeover process 200 continues to an
13 'elective takeover Save State' procedure' 223.

14
15 An 'elective reboot save state' procedure 219 causes the current state of the
16 first server 140 to be stored in the first non-volatile storage 141. This includes the setting
17 of the flag value to indicate a planned reboot of the first server 140.

18
19 A 'shut down system' procedure 221 causes the first server 140 to be shut
20 down. The file server elective reboot/takeover process 200 terminates through an "end"
21 terminal 229.

22
23 An 'elective takeover save state' procedure 223 causes the current state of
24 the first server 140 to be stored in the first non-volatile storage 141 and the second non-
25 volatile storage 151.

1 A 'takeover server restore state' procedure 225 allows the state of the first
2 server 140 stored in the first non-volatile storage 141 to be transferred via the
3 interconnect 170 and reconstituted on the second server device 150 or procured from the
4 second non-volatile storage 151. At this point the second server 150 is supporting the
5 sessions that were active on the first server 140 prior to elective takeover and CIFS
6 processing within these sessions continues. The file server elective reboot/takeover
7 process 200 terminates through an "end" terminal 229.

8
9 A 'restore state' procedure 227 allows the state of the first server 140 to be
10 reconstituted to the state it was in prior to an elective reboot or non-elective reboot from
11 the state stored in the first non-volatile storage 141. The file server elective
12 reboot/takeover process 200 continues to a 'receive CIFS requests' procedure 207.

13
14
15 *Method of Operation – Non-Elective Reboot.*

16
17 Figure 3 illustrates a file server non-elective reboot process, indicated by
18 general reference character 300. The file server non-elective reboot process 300 initiates
19 at a 'start' terminal 301. The file server non-elective reboot process 300 continues to a
20 'boot system' procedure 303 which allows the first server 140 to boot.

21
22 A "flag active" decision procedure 305 determines whether an non-elective
23 reboot has occurred. If the 'flag active' decision procedure 305 determines that the a
24 non-elective reboot has occurred, the file server non-elective reboot process 300
25 continues to the 'resume normal operation' procedure 309.

1 A 'restore state' procedure 307 allows the first server 140 to reconstitute
2 the state it was in prior to the non-elective reboot by copying state from that stored in the
3 first non-volatile storage 141 or second non-volatile storage 151.

4
5 A 'resume normal operation' procedure 309 allows the first server 140 to
6 once again accept and process CIFS requests and perform all functions it was executing
7 prior to the non-elective reboot.

8
9 The file server non-elective reboot process 300 terminates through an "end"
10 terminal 311.

11
12
13 *Method of Operation – Non-Elective Takeover.*

14
15
16 Figure 4 illustrates a file server non-elective takeover process, indicated by
17 general reference character 4. The file server non-elective takeover process 400 initiates
18 at a 'start' terminal 401. The file server non-elective takeover process 400 continues to a
19 'boot system' procedure 403 which allows the first server 140 to boot.

20
21 A 'receive CIFS requests' procedure 405 allows user requests to be
22 received by the first server 140.

23
24 A 'process CIFS requests' procedure 407 allows the first server 140 to
25 respond to requests from the client device 110 by providing access to data contained in
26 the mass storage 180.

1
2 A 'save state' procedure 409 allows the state of the first server 140 to be
3 saved to the first non-volatile storage 141 and the second non-volatile storage 151. In a
4 preferred embodiment, reliably state saving in anticipation of a system failure may be
5 performed at any one of a plurality of specific points within the processing of CIFS
6 requests. For clarity in the description of this method of operation, 'save state' procedure
7 409 is indicated only once. The specific points for saving state are further discussed
8 within this application.

9
10 A 'filer1 failure' decision procedure 411 determines whether the first server
11 140 has failed in some way. In a preferred embodiment, failure of the first server 140
12 would be detected by the second server 150. If the 'filer1 failure' decision procedure 411
13 determines that the first server 140 has not failed, the file server non-elective takeover
14 process 400 continues to the 'receive CIFS requests' procedure 405.

15
16 A 'restore state' procedure 413 allows the state of the First server 140 prior
17 to failure to be reconstitute on the second server 150 by copying state from that stored in
18 the first non-volatile storage 141 or second non-volatile storage 151.

19
20 A 'filer2 takeover' procedure 415 completes the process by allowing the
21 second filer 150 to resume processing of CIFS request where the first server 140 stopped.

22
23 The file server non-elective takeover process 400 terminates through an
24 "end" terminal 417.

1 *Method of Operation – Automatic State Saving*

2
3 Figure 5 illustrates critical state saving points in a mechanism to survive
4 server failures when using the CIFS protocol.

5
6 The saved state must always be in a consistent state. Automatic state
7 saving must occur at specified points within a session of communication between the first
8 server 140 and the client device 110 to ensure that the saved state is consistent.

9
10 POINT 1: State is saved prior to TCP acknowledging an incoming CIFS request.
11 If the system fails prior to this, then the effect is as if the packet was never received, and
12 retransmission by the client device 110 occurs. If the system fails after the
13 acknowledgment is sent, then the system has a record that the request came in and it will
14 be processed when state is restored.

15
16 POINT 2: State is saved prior to CIFS starting a SMB command. If the system
17 fails prior to this, TCP will redeliver the TCP message to CIFS. If the system fails after
18 this, the saved state indicates that the first server 140 started work on a CIFS operation.
19 (Some single CIFS commands are composite operations: e.g. open, read, and close. In
20 such cases, saving state is required before each component operation).

21
22 POINT 3: State is saved when a CIFS operation completes. If the system fails
23 prior to this, the same CIFS operation is repeated creating the same result. If the system
24 fails after this, the reply is sent again and TCP treats it as a duplicate.

1 POINT 4: State is saved after TCP acknowledges the reply. If the system
2 fails prior to this, then the reply never happened and will be sent again. If the system
3 fails after the acknowledgment but before the acknowledgment is saved, then we will
4 duplicate the acknowledgment and normal TCP handling will process that without any
5 problems. If the system fails after the save has occurred the acknowledgment will not be
6 repeated.

7
8 These four points illustrate where state may be saved in a consistent
9 manner, however, there are other points where state may be reliably saved and these
10 points would be obvious to one skilled in the art.

11
12
13 *Generality of the Invention*

14
15 The invention has general applicability to various fields of use, not necessarily
16 related to the services described above. For example, these fields of use can include one
17 or more of, or some combination of, the following:

- 18
19 • In addition to general applicability to CIFS the invention has broad applicability to
20 other transmission protocols.

21
22 Other and further applications of the invention, in its most general form, will be clear
23 to those skilled in the art after perusal of this application, and are within the scope and
24 spirit of the invention.

